# 2006-2017年我国外语测试实证研究：回顾与展望

江进林

对外经济贸易大学

# 研究背景

- 学界已对1996-2005年国内的外语测试研究进行了回顾（蒋显菊 2007），但近十余年的研究进展尚缺乏梳理。
- 相当一部分研究是书评、理论介绍、大纲或测试方法的操作描述、现状反思类文章（Cheng 2008），实证研究有待进一步发展。

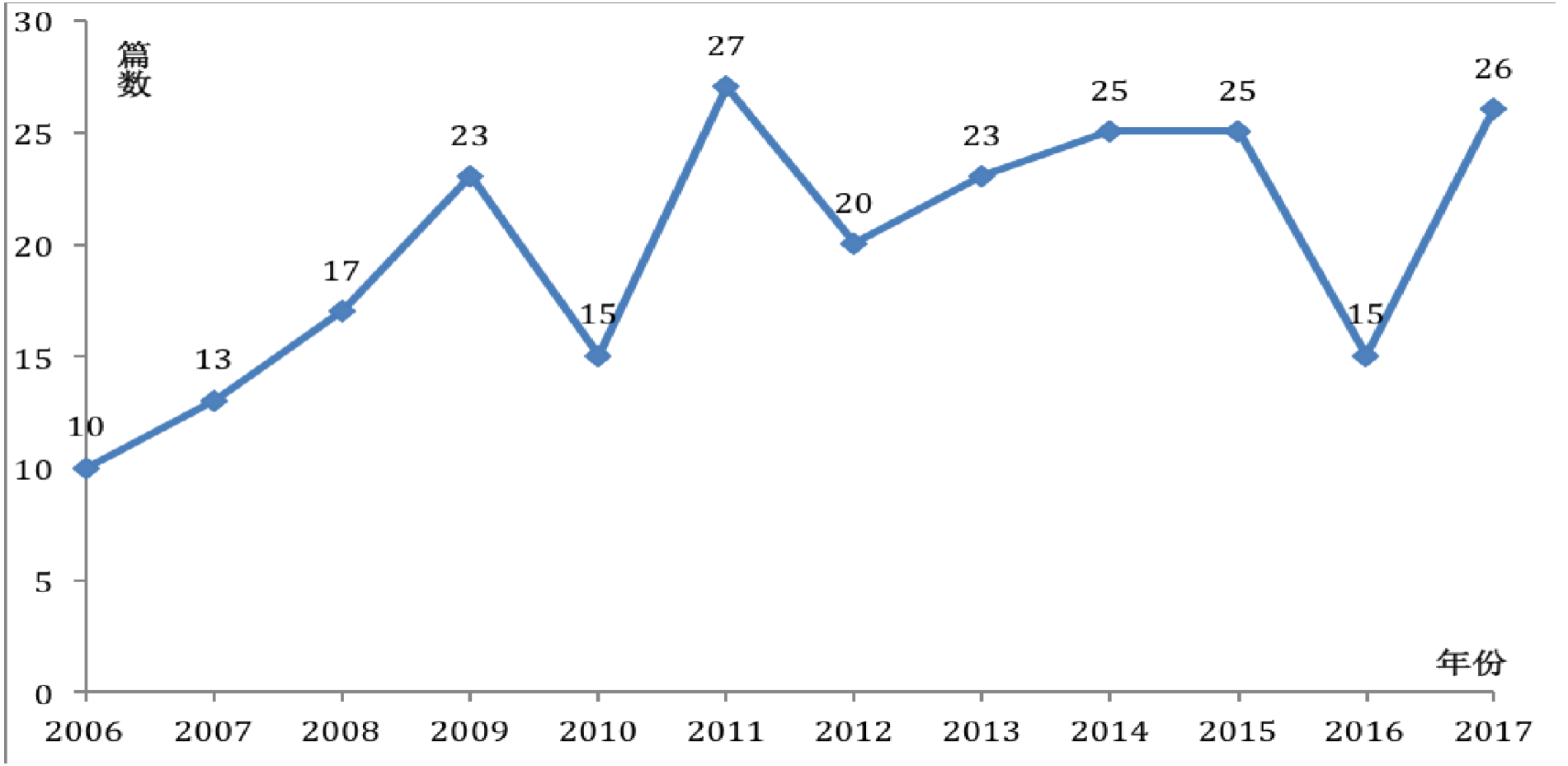
# 研究目的

- 检索整理我国外语类主要期刊2006-2017年发表的外语测试实证研究，分析研究的整体趋势、内容和特点。

# 研究设计

- 中国知网
- 2006-2017年12种外语类主要期刊（11种CSSCI+《外语测试与教学》）
- 内容分析法，略读题目、摘要、论文
- 239篇外语测试相关的实证研究论文

# 年份分布


30
25
20
15
10
5
0
篇数
10
13
17
23
15
27
20
23
25
25
15
26
年份
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

# 年份分布

- 波动式发展的趋势
- 2009、2011和2017年达到三个峰值

# 年份分布

- 2009年发表实证研究最多的是《外语界》（8篇），其中6篇与机考和网考相关，因为CET4&6从2007年开始试点网考。
- 2011年《外语测试与教学》创刊并发表了16篇实证研究论文。
- 2017年发表最多的是《外语界》和《外语测试与教学》（各10篇），其中7篇与中国英语能力等级量表有关。

# 年份分布

容易成为学界关注的焦点：

- 大型考试的改革
- 外语测试体系的改革

# 测试分布

表2　外语测试实证研究关注的测试类型

| 测试类型 | 篇数 | 百分比 |
| --- | --- | --- |
| 高考 | 11 | 4.60% |
| 大学英语四、六级 | 53 | 22.18% |
| 英语专业四、八级 | 40 | 16.74% |
| BEC、PETS、雅思、托福、PISA、研究生入学英语考试、口译岗位资格证书考试、商务英语专业四级考试等 | 20 | 8.37% |
| 课程评估、校本考试、自主开发的测试 | 115 | 48.12% |

# 测试分布

- 超过一半（51.89%）的实证研究关注大规模测试。大学英语四、六级（22.18%）和英语专业四、八级考试（16.74%）的关注度较高，BEC、PETS、雅思、托福等其他大型考试（8.37%）和高考（4.60%）的相关研究占比较低。
- 近一半研究（48.12%）关注课程评估、校本考试或自主开发的测试。

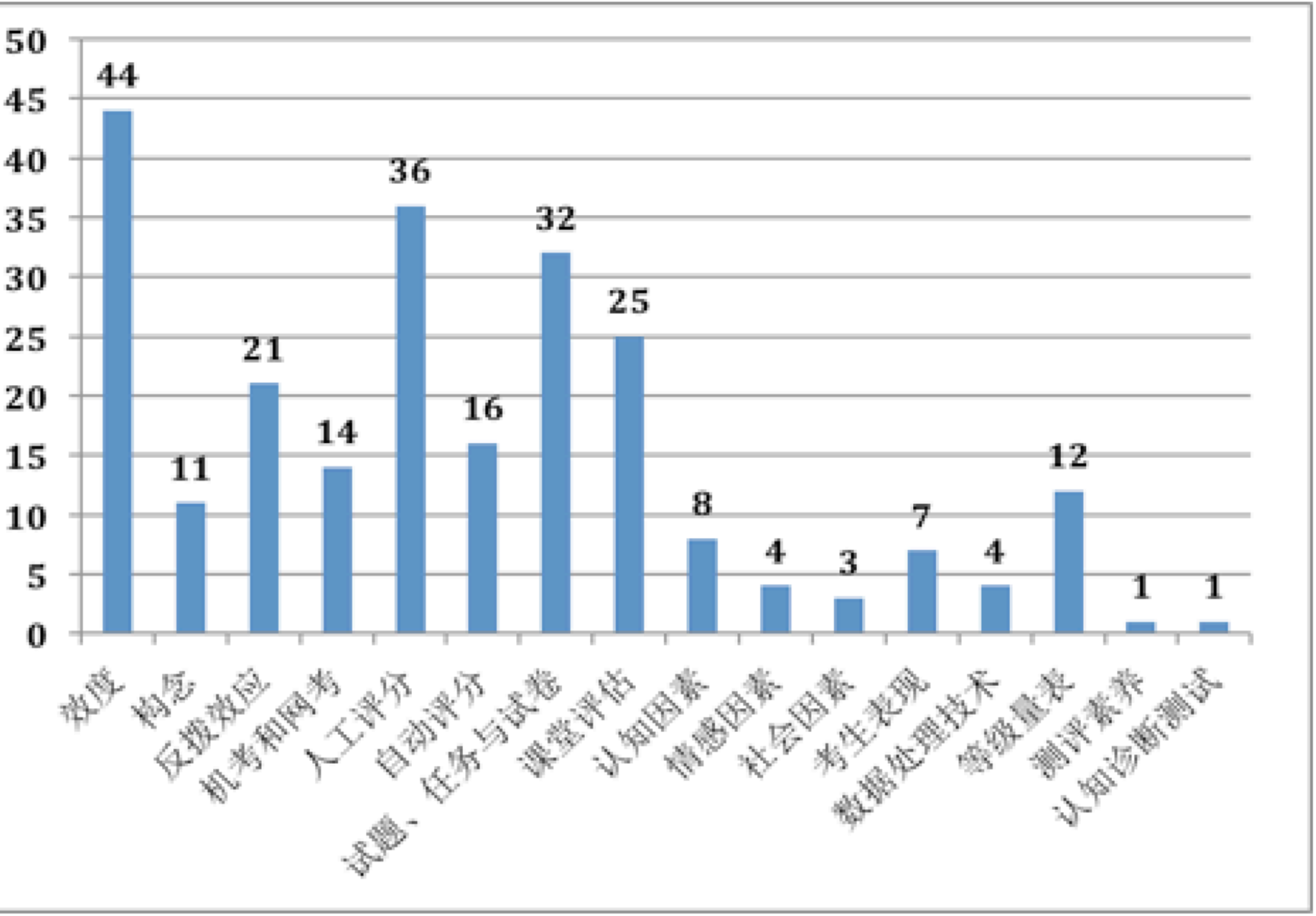



图 2　外语测试实证研究的主题分布

# 主题分布

- 内容大体涵盖16个主题：
- 效度研究最多（44篇，占18.41%）；
- 其次是人工评分研究（36篇，占15.06%）；
- 排第三位的是试题、任务与试卷研究（32篇，占13.39%）；
- 第四和第五位分别是课堂评估研究（25篇，占10.46%）和反拨效应研究（21篇，占8.79%）；
- 认知因素、情感因素、社会因素、考生表现、数据处理技术受到的关注较少，均不到10篇；
- 测评素养和认知诊断测试研究最少，各1篇。

# 主题分布

- 第1位：效度研究，主要关注专四和专八考试。
- 专四效度验证的对象较多：口试、语法词汇项目、阅读理解和完形填空；效度证据的来源也比较广泛。
- 口试的效度证据包括与教学大纲的内容相关性、与笔试成绩的标准关联性、测试内容的长度和难度、施测过程和评分、表面效度、预示效度、共时效度、理论效度；
- 词汇语法项目的效度检验包括内容效度、基于Rasch模型的答题反应分析和基于因子分析的构念效度；
- 阅读理解和完形填空的效度证据主要来自答题过程和项目功能差异分析。

# 主题分布

- 专八的效度验证对象主要是人文知识项目：
- （1）内容效度，如内容关联性和覆盖面；
- （2）统计属性，如难度、区分度。

# 主题分布

- 第2位：人工评分研究集中于四个方面。
- 评分方法。比较分项评分法、整体评分法及其变体形式在作文、口语、翻译评价中的应用，发现两者各有利弊。
- 评分信度。采用多面Rasch模型从试题、考生、评分员、评分量表等方面对口试、作文、翻译的分数进行了分析。
- 评分员效应。采用多面Rasch模型探讨了评分员主观因素对口试、作文、翻译分数的影响，如严厉度、趋中性、集中趋势、随机效应、晕轮效应。少数研究探讨了评分员在理解和使用评分标准上的差异。
- 评分员培训。教师在培训前后与专家的一致性，发现培训提高了口试评分的准确度。

# 主题分布

- 第3位：试题、任务与试卷
- （1）分析试题、任务与试卷的特征；
- （2）考察不同任务和题型对成绩或答题行为的影响，包括：
- 听力题干的输入模式、不同的听写测试方法、视听试题设计和音视频特征；
- 阅读的语篇类型和题型、生词密度、单项选择题的排序；
- 完形填空的语篇类型、删词类型和答题方法；
- 写作题目、提示句、图表等信息。

# 主题分布

- 第4位：课堂评估
- 研究者将同伴评价、形成性评价、动态评估等应用于写作、阅读、翻译、口语等教学，取得了良好的效果。

# 主题分布

- 第5位：反拨效应
- 研究对象集中于四、六级考试，研究方法主要是问卷调查和访谈，调查群体基本上是学生，调查的内容包括考试对其学习内容和方式、深度和广度等方面的影响。
- 少数研究者也对专四、专八考试的反拨作用进行了调查，调查的对象是外语专家、教师和学生。
- 研究发现，这些考试的正面反拨作用大于负面效应。

# 主题分布

- 第6位：自动评分
- 作文自动评分研究日趋成熟，已开始应用于写作教学。
- 少数研究者研发了能够对翻译进行自动评分的模型，模型对译文语义内容的评分与人工评分的相关度在0.8以上，对语言形式评分的相关度在0.6以上（江进林 2013；江进林、文秋芳 2010）。
- 口语自动评分研究开始起步，朗读题型的机器阅卷与人工评分的相关度达到0.713（李萌涛等 2008）。

# 主题分布

- 第7位：机考和网考包括三个方面。
- 考试的信、效度。研究者比较了电脑写作和纸笔写作、机考口试和面试型口试的成绩，发现机考与传统考试的信、效度基本一致。不过，机考会对口语产出产生影响。在互动性的口语任务中，学生错误更多且语速较慢。
- 考生的态度。已有研究从机考过程观察、师生反馈、题型分析、考生焦虑度调查等多个角度论证了机考的优缺点。
- 人工评分。研究者通过比较发现作文网上阅卷的信度高于传统阅卷。

# 主题分布

- 第8位：等级量表
- 中国英语能力等级量表开发近两年已取得实质性进展。
- 研究者分析了量表的制定方法，探究了听力、口语、阅读、写作描述语的开发和验证过程（如何莲珍，陈大建 2017；曾用强 2017）。

# 主题分布

- 第9位：构念研究探讨了写作、口译、朗读、听力理解、语用等外语能力要素的构成。
- 研究方法包括对专家和评分员的问卷调查、因子分析等。
- 如：研究发现，英语专业学生的写作能力构念包括思想表达、组织结构、语言的准确性、丰富性、得体性、写作规范等。

# 主题分布

- 第10位：认知因素：考生的策略使用。
- 第11位：情感因素：集中于学习动机、焦虑及其对成绩或考试表现的影响。
- 第12位：社会因素：少数研究开始关注社会因素对测试的影响，例如，周季鸣和刘琨（2011）考察了测试结果的使用政策如何影响师生对测试的认识、准备和测试结果。
- 第13位：考生表现：分析考生在测试中反映出的问题和困难，对教学提出相关建议。
- 第14位：数据处理技术：试卷等值、分数等值、垂直等值技术，以及Rasch模型的应用。

# 主题分布

- 第15位：测评素养：唐雄英（2017）通过分析测评课程学员的学习日志等资料，探索了在职中小学英语教师测评素养的发展途径。
- 第16位：认知诊断测试：陈慧麟和赵冠芳（2013）运用认知诊断测试对PISA英语阅读测试的结果进行了实证分析，验证了对语言测试进行认知诊断的可操作性。

# 研究特点

- （1）研究对象广泛
- 既涵盖各级各类大规模高风险测试，也包括课程评估、校本考试、自主开发的测试。
- 研究的对象涉及不同层次与类型的学生、教师和其他利益相关者，如外语专家、考试设计者和命题人员。

# 研究特点

- （2）研究内容多元
- 涵盖效度、构念、反拨效应、人工评分、试题/任务/试卷、认知因素、情感因素、考生表现等传统研究主题；
- 也包括机考和网考、自动评分、课堂评估、等级量表、测评素养、认知诊断测试等改革举措。
- 更重要的是，学生掌握了一部分自主评价的权力，同伴评价、形成性评价等已初步应用于课堂教学。可见国内研究者越来越深入地认识到测试与教学的关系。这与1996-2005年国内的研究（蒋显菊 2007）相比是很大的进步。

# 研究特点

- （3）研究方法综合
- 多数研究结合使用量化与质性研究方法。
- 具体而言，已有研究主要采用6种方法收集数据：问卷调查、访谈、测试、考试过程观察、课堂观察、反思日志。
- 运用最多的是问卷调查和访谈。
- 很多研究也使用量化统计和测量技术来处理数据，包括多面Rasch模型、因子分析、概化理论、结构方程模型等。

# 研究特点

- 不足：
- （1）研究对象主要集中于本科生，少量研究关注教师和其他利益相关者。鲜有研究涉及中小学生、专科生、研究生等学生群体。

# 研究特点

- （2）有些考试的研究内容需要进一步拓展深化。如：
- 绝大多数效度研究集中于专四和专八考试，针对高考的研究仅3项；
- 反拨效应研究的对象主要是四、六级考试，针对高考和专四、专八考试的都仅寥寥几项；
- 对口试任务类型的研究也比较欠缺。

# 研究特点

- （3）研究主题的前沿性有待提升，研究成果的实用性与国际研究相比较为欠缺。
- 国外近五年研究热点：效度、人工评分、专门用途英语测试、认知诊断测试、测评素养
- 对如何改进专门用途英语测试、如何以评促学、如何对教师进行测评素养培训提供了启示。
- 国内近年开始出现相关研究，但实证研究极少。

# 总结

- 需要关注中小学生、专科生、研究生等学生群体；
- 拓展研究较少的领域，如口试任务类型、高考的效度和反拨效应；
- 关注专门用途英语测试、认知诊断测试、测评素养等国际研究前沿，并将研究成果用于测试开发、课堂教学和教师教育课程，提高实用性。

谢谢!